Springboard DSC Program

Capstone Project 2

Using Eye Movement during Reading to predict Dyslexia

Screening for Dyslexia Using Eye Tracking During Reading

By Harinee Madhusudhan

June 2020

Table of Contents

Introduction

Dyslexia is a neurodevelopmental disability that adversely affects the speed and accuracy of word recognition, and therefore, impedes reading fluency and text comprehension. It is commonly estimated to affect between 5 and 10 percent of the population.

Eye tracking has been increasing in popularity over the past decade as a window into observers' visual and cognitive processes. For instance, researchers have utilized eye tracking to study behavior in such domains as image scanning, driving, arithmetic, analogy, and reading. The process of fixation identification—separating and labeling fixations and saccades in eye-tracking protocols—is an essential part of eye-movement data analysis and can be used for predicting Dyslexia.

The implementation details can be found in the notebooks I developed, which at <https://github.com/harineem1/Springboard01/tree/master/Capstone/Capstone%202>

Approach

Data Acquisition and Wrangling

The experiments report is based on eye tracking data from 185 subjects participating in the longitudinal research project on reading development and reading disability in Swedish school children running between 1989 and 2010.The inclusion criteria required that subjects (1) had Swedish as first language; (2) performed in the lower 5th percentile of the full cohort on two standardized tests of word decoding; and, (3) experienced persistent problems in learning to read according to an independent assessment completed by the classroom teacher.

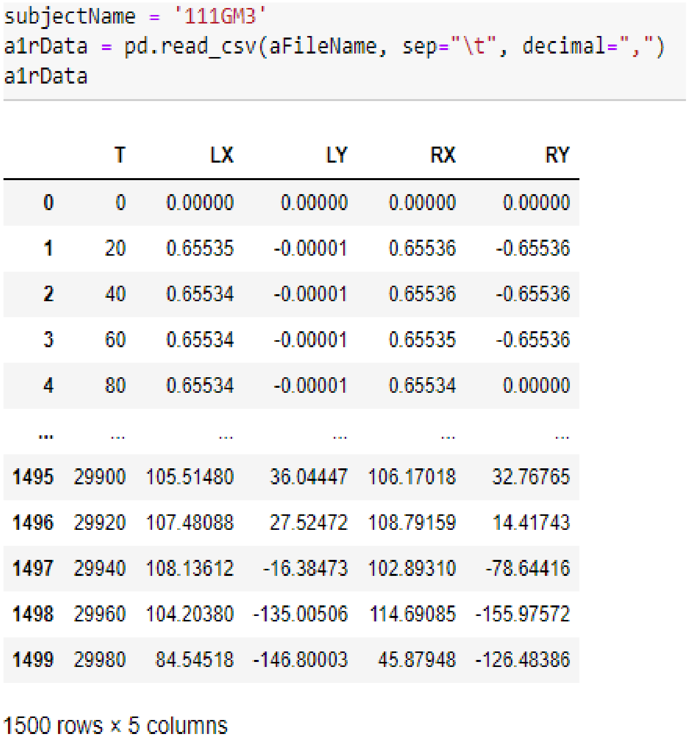
It is worth noting, however, that we do not know how many of the original HR subjects received an actual diagnosis of dyslexia later. The main reason for this is that during the initial years of data collection the notion of dyslexia was still not well established in pedagogic practices in Sweden and very few individuals were diagnosed in general. As dyslexia diagnoses became more common over the years, most of the HR subjects in the study had already finished school which further reduced their likelihood of receiving a diagnosis.

While the subjects were attending 3rd grade (age 9–10), eye movements were recorded as part of an ophthalmological examination that aimed to investigate whether there were any differences between the two groups in terms of basic visual and oculomotor functions. While some minor differences were reported, it was concluded that these differences most likely reflected secondary effects of the cognitive difficulties that the HR subjects experienced with language processing, rather than inherent visual deficits. For the present experiment, we use eye movement recordings made while the subjects were reading a short natural passage of text adapted to their age. Recordings were available for 185 subjects, 97 High Risk subjects (76 males and 21 females) and 88 Low Risk subjects (69 males and 19 females).

A goggle-based infrared corneal reflection system was used to track eye position over time, sampling the horizontal and vertical position of both eyes. During recording, subjects were equipped with a pair of lightweight (80g), individually adjustable, head-mounted goggles in which four arrays of infrared transmitters and detectors were mounted, arranged in a square around each eye. A chin and forehead rest were deployed to minimize head movements and stabilize the viewing distance at 45 cm. Since calibration was done manually, the gain for horizontal movements of the left eye was first set, then the gain for horizontal movements of the right eye and so on for vertical movements

All subjects read one and the same text presented on a single page of white paper with high contrast. The text was distributed over 8 lines and consisted of 10 sentences with an average length of 4.6 words.

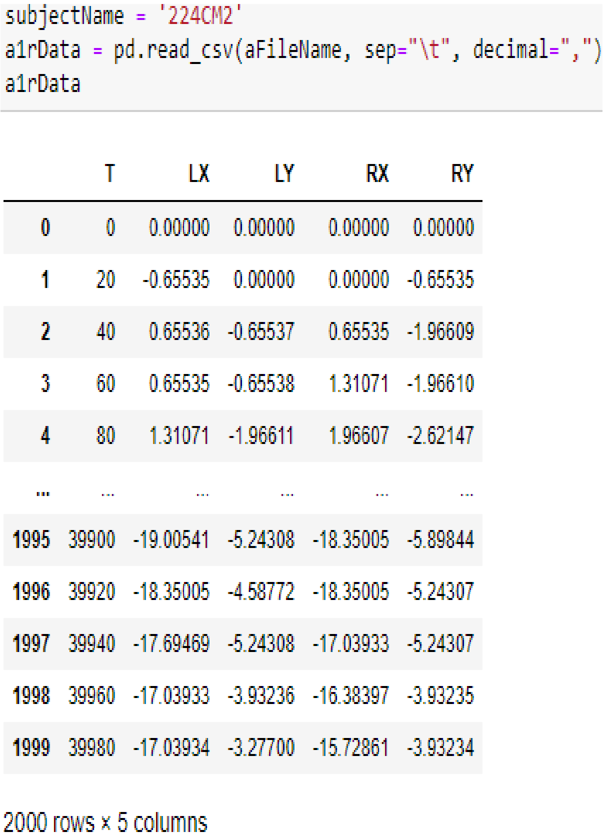
The data consists of 185 folders for each subject and has left and right eye movement with time interval of 20 microseconds. The number of rows of each subject data are different as according to the reading ability of the subject.



```
subjectName = '111GM3'
a1rData = pd.read_csv(aFileName, sep="\t", decimal=",")
a1rData
```

	T	LX	LY	RX	RY
0	0	0.00000	0.00000	0.00000	0.00000
1	20	0.65535	-0.00001	0.65536	-0.65536
2	40	0.65534	-0.00001	0.65536	-0.65536
3	60	0.65534	-0.00001	0.65535	-0.65536
4	80	0.65534	-0.00001	0.65534	0.00000
...	...	...	...	...	...
1495	29900	105.51480	36.04447	106.17018	32.76765
1496	29920	107.48088	27.52472	108.79159	14.41743
1497	29940	108.13612	-16.38473	102.89310	-78.64416
1498	29960	104.20380	-135.00506	114.69085	-155.97572
1499	29980	84.54518	-146.80003	45.87948	-126.48386

1500 rows × 5 columns



```
subjectName = '224CM2'
a1rData = pd.read_csv(aFileName, sep="\t", decimal=",")
a1rData
```

	T	LX	LY	RX	RY
0	0	0.00000	0.00000	0.00000	0.00000
1	20	-0.65535	0.00000	0.00000	-0.65535
2	40	0.65536	-0.65537	0.65535	-1.96609
3	60	0.65535	-0.65538	1.31071	-1.96610
4	80	1.31071	-1.96611	1.96607	-2.62147
...	...	...	...	...	...
1995	39900	-19.00541	-5.24308	-18.35005	-5.89844
1996	39920	-18.35005	-4.58772	-18.35005	-5.24307
1997	39940	-17.69469	-5.24308	-17.03933	-5.24307
1998	39960	-17.03933	-3.93236	-16.38397	-3.93235
1999	39980	-17.03934	-3.27700	-15.72861	-3.93234

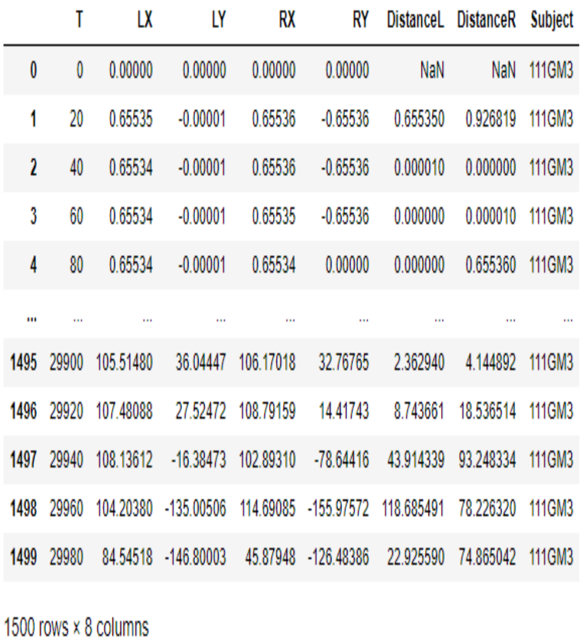
2000 rows × 5 columns

Subjects with a code ending in 1 or 2 were reading disabled. Subjects with a code ending in 3 or 4 were controls. Subjects with a code ending in 1 or 3 were male. Subjects with a code ending in 2 or 4 were female.

In the above example, subject 111GM3 was a male in the control group. The subject 224CM2 was a female diagnosed as dyslexic.

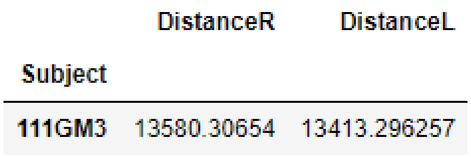
Extracting Features.

Using the eye movement DistanceL and DistanceR are computed. The DistanceL and DistanceR were computed as the distance traveled in the time interval using the formula

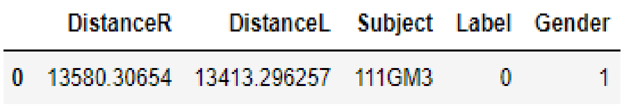


	T	LX	LY	RX	RY	DistanceL	DistanceR	Subject
0	0	0.00000	0.00000	0.00000	0.00000	NaN	NaN	111GM3
1	20	0.65535	-0.00001	0.65536	-0.65536	0.655350	0.926819	111GM3
2	40	0.65534	-0.00001	0.65536	-0.65536	0.000010	0.000000	111GM3
3	60	0.65534	-0.00001	0.65535	-0.65536	0.000000	0.000010	111GM3
4	80	0.65534	-0.00001	0.65534	0.00000	0.000000	0.655360	111GM3
...	...	...	...	...	...	...	...	...
1495	29900	105.51480	36.04447	106.17018	32.76765	2.362940	4.144892	111GM3
1496	29920	107.48088	27.52472	108.79159	14.41743	8.743661	18.536514	111GM3
1497	29940	108.13612	-16.38473	102.89310	-78.64416	43.914339	93.248334	111GM3
1498	29960	104.20380	-135.00506	114.69085	-155.97572	118.685491	78.226320	111GM3
1499	29980	84.54518	-146.80003	45.87948	-126.48386	22.925590	74.865042	111GM3

1500 rows × 8 columns



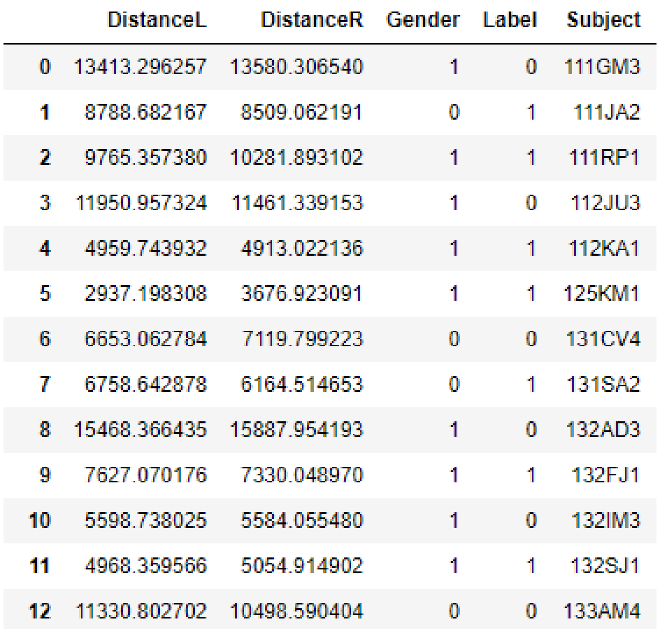
	DistanceR	DistanceL
Subject		
111GM3	13580.30654	13413.296257



	DistanceR	DistanceL	Subject	Label	Gender
0	13580.30654	13413.296257	111GM3	0	1

DistanceL = sqrt[(LX -LX’)**2 + (LY-LY’)**2] where (LX, LY) is the first location of the left eye and (LX’, LY’) is the next location of the left eye. Similarly, the DistanceR was computed using the data RX and RY.

The sum of the distances travelled provides an idea of the total distance traveled for each eye. This value is computed for each subject. A new data frame is created with this distance traveled data and the columns Subject, Label and Gender.

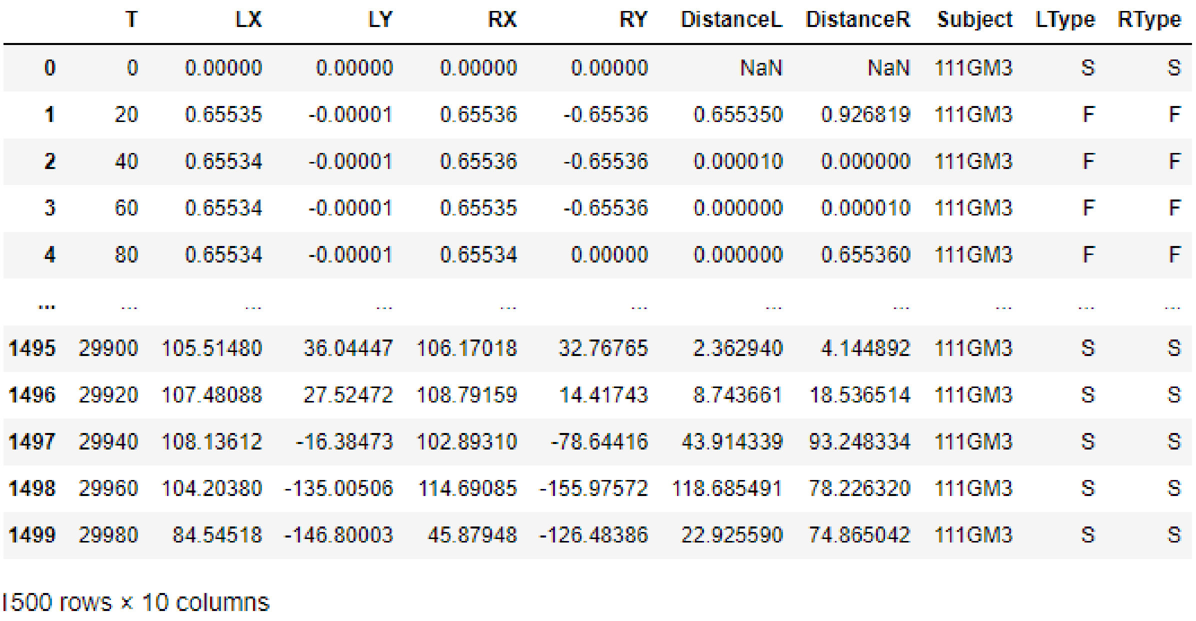


	DistanceL	DistanceR	Gender	Label	Subject
0	13413.296257	13580.306540	1	0	111GM3
1	8788.682167	8509.062191	0	1	111JA2
2	9765.357380	10281.893102	1	1	111RP1
3	11950.957324	11461.339153	1	0	112JU3
4	4959.743932	4913.022136	1	1	112KA1
5	2937.198308	3676.923091	1	1	125KM1
6	6653.062784	7119.799223	0	0	131CV4
7	6758.642878	6164.514653	0	1	131SA2
8	15468.366435	15887.954193	1	0	132AD3
9	7627.070176	7330.048970	1	1	132FJ1
10	5598.738025	5584.055480	1	0	132IM3
11	4968.359566	5054.914902	1	1	132SJ1
12	11330.802702	10498.590404	0	0	133AM4

Go back to a1rData and extract few more columns from the eye movement data: We need to distinguish between the two movement types, Fixation and Saccade. A fixation is when the eyes do not move and are focused at one location. A saccade is the movement of the eye from one location to another.

We add to columns LType and RType, which could contain one of two values, F or S. If the distance traveled from the last time is less than 1 unit, then we consider the eye to be focused at one location. If the distance if > 1 unit then we assume that the eyes are moving and hence they are in a saccade state. For each observation we compute if the eyes are in fixation or saccade state.

S Saccade F Fixation: If the distance travelled is < 1 then it is Fixation else Saccade



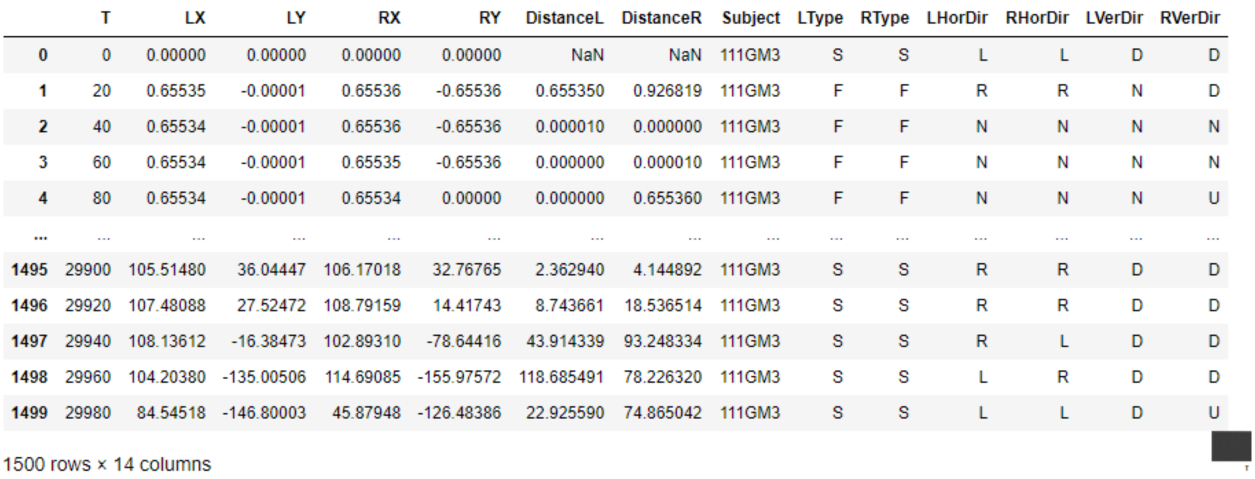
	T	LX	LY	RX	RY	DistanceL	DistanceR	Subject	LType	RType
0	0	0.00000	0.00000	0.00000	0.00000	NaN	NaN	111GM3	S	S
1	20	0.65535	-0.00001	0.65536	-0.65536	0.655350	0.926819	111GM3	F	F
2	40	0.65534	-0.00001	0.65536	-0.65536	0.000010	0.000000	111GM3	F	F
3	60	0.65534	-0.00001	0.65535	-0.65536	0.000000	0.000010	111GM3	F	F
4	80	0.65534	-0.00001	0.65534	0.00000	0.000000	0.655360	111GM3	F	F
...	...	...	...	...	...	...	...	...	...	...
1495	29900	105.51480	36.04447	106.17018	32.76765	2.362940	4.144892	111GM3	S	S
1496	29920	107.48088	27.52472	108.79159	14.41743	8.743661	18.536514	111GM3	S	S
1497	29940	108.13612	-16.38473	102.89310	-78.64416	43.914339	93.248334	111GM3	S	S
1498	29960	104.20380	-135.00506	114.69085	-155.97572	118.685491	78.226320	111GM3	S	S
1499	29980	84.54518	-146.80003	45.87948	-126.48386	22.925590	74.865042	111GM3	S	S

1500 rows × 10 columns

Similarly, see whether the movement is right/left for horizontal or (up/down) for vertical for each of the eye (left & right)

-1 left/Down 1 Right/Up 0 No change: Identify the Left and Right horizontal and vertical direction

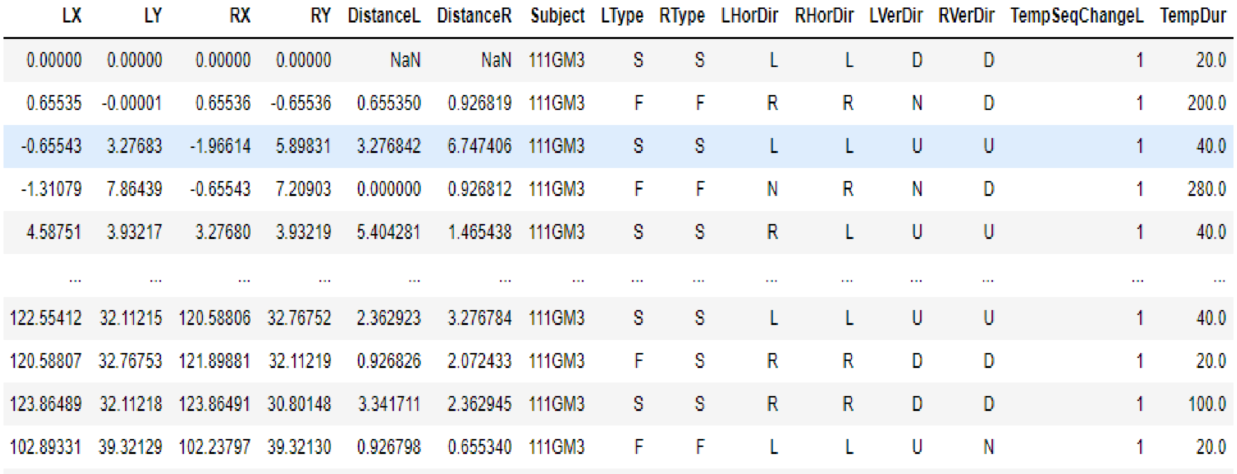
LHorDir	the movement of Left eye in the horizontal direction	L = Left, R = Right, N = No Change
RHorDir	the movement of Right eye in the horizontal direction	L = Left, R = Right, N = No Change
LVerDir	the movement of Left eye in the vertical direction	U = Up, D = Down, N = No Change
RVerDir	the movement of Right eye in the vertical direction	U = Up, D = Down, N = No Change



	T	LX	LY	RX	RY	DistanceL	DistanceR	Subject	LType	RType	LHorDir	RHorDir	LVerDir	RVerDir
0	0	0.00000	0.00000	0.00000	0.00000	NaN	NaN	111GM3	S	S	L	L	D	D
1	20	0.65535	-0.00001	0.65536	-0.65536	0.655350	0.926819	111GM3	F	F	R	R	N	D
2	40	0.65534	-0.00001	0.65536	-0.65536	0.000010	0.000000	111GM3	F	F	N	N	N	N
3	60	0.65534	-0.00001	0.65535	-0.65536	0.000000	0.000010	111GM3	F	F	N	N	N	N
4	80	0.65534	-0.00001	0.65534	0.00000	0.000000	0.655360	111GM3	F	F	N	N	N	U
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1495	29900	105.51480	36.04447	106.17018	32.76765	2.362940	4.144892	111GM3	S	S	R	R	D	D
1496	29920	107.48088	27.52472	108.79159	14.41743	8.743661	18.536514	111GM3	S	S	R	R	D	D
1497	29940	108.13612	-16.38473	102.89310	-78.64416	43.914339	93.248334	111GM3	S	S	R	L	D	D
1498	29960	104.20380	-135.00506	114.69085	-155.97572	118.685491	78.226320	111GM3	S	S	L	R	D	D
1499	29980	84.54518	-146.80003	45.87948	-126.48386	22.925590	74.865042	111GM3	S	S	L	L	D	U

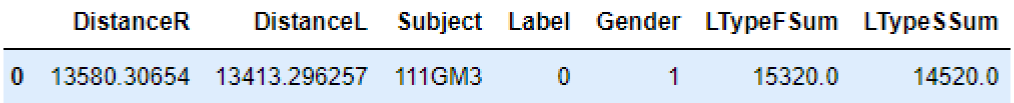
1500 rows × 14 columns

Now we need to find the time taken for each of the movements, fixation, saccade, Ups, Downs, Lefts Rights, No changes etc. For each of these, create a temp column that notes if there is a change in them from the last time. For example, Row 0 and 1 were in S state. Row 2 – 9 are in F state. We aggregate the sum of fixations and saccades, the counts, and their mean times. A fixation state is detected when the eyes have remained stable between wo measurements, and a saccade state when the eyes have moved beyond the threshold distance (1 unit). Once a change of state is detected, the samples of the previous state are identified as a new event using the data item TempSeqChange.



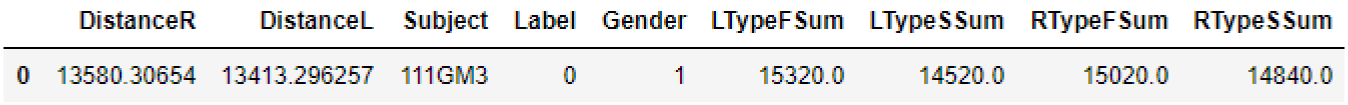
LX	LY	RX	RY	DistanceL	DistanceR	Subject	LType	RType	LHorDir	RHorDir	LVerDir	RVerDir	TempSeqChangeL	TempDur
0.00000	0.00000	0.00000	0.00000	NaN	NaN	111GM3	S	S	L	L	D	D	1	20.0
0.65535	-0.00001	0.65536	-0.65536	0.655350	0.926819	111GM3	F	F	R	R	N	D	1	200.0
-0.65543	3.27683	-1.96614	5.89831	3.276842	6.747406	111GM3	S	S	L	L	U	U	1	40.0
-1.31079	7.86439	-0.65543	7.20903	0.000000	0.926812	111GM3	F	F	N	R	N	D	1	280.0
4.58751	3.93217	3.27680	3.93219	5.404281	1.465438	111GM3	S	S	R	L	U	U	1	40.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
122.55412	32.11215	120.58806	32.76752	2.362923	3.276784	111GM3	S	S	L	L	U	U	1	40.0
120.58807	32.76753	121.89881	32.11219	0.926826	2.072433	111GM3	F	S	R	R	D	D	1	20.0
123.86489	32.11218	123.86491	30.80148	3.341711	2.362945	111GM3	S	S	R	R	D	D	1	100.0
102.89331	39.32129	102.23797	39.32130	0.926798	0.655340	111GM3	F	F	L	L	U	N	1	20.0

We compute each total Fixation duration and Saccade duration by summing up the time in each of the events.



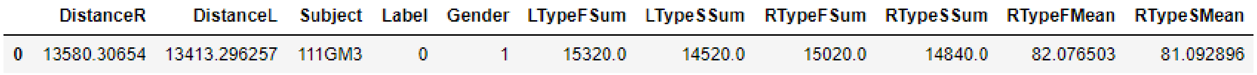
	DistanceR	DistanceL	Subject	Label	Gender	LTypeFSum	LTypeSSum
0	13580.30654	13413.296257	111GM3	0	1	15320.0	14520.0

And repeat the same for RType.

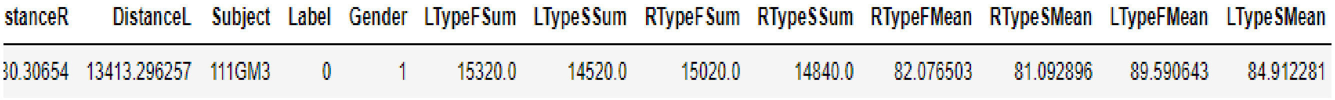


	DistanceR	DistanceL	Subject	Label	Gender	LTypeFSum	LTypeSSum	RTypeFSum	RTypeSSum
0	13580.30654	13413.296257	111GM3	0	1	15320.0	14520.0	15020.0	14840.0

compute the mean of F durations and S durations, grouping by RType and LType

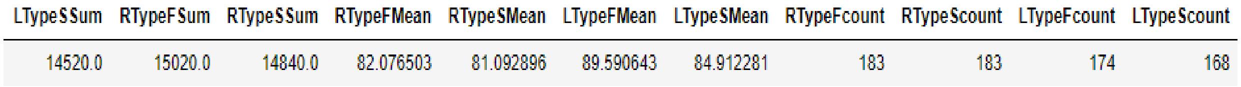


	DistanceR	DistanceL	Subject	Label	Gender	LTypeFSum	LTypeSSum	RTypeFSum	RTypeSSum	RTypeFMean	RTypeSMean
0	13580.30654	13413.296257	111GM3	0	1	15320.0	14520.0	15020.0	14840.0	82.076503	81.092896



stanceR	DistanceL	Subject	Label	Gender	LTypeFSum	LTypeSSum	RTypeFSum	RTypeSSum	RTypeFMean	RTypeSMean	LTypeFMean	LTypeSMean
30.30654	13413.296257	111GM3	0	1	15320.0	14520.0	15020.0	14840.0	82.076503	81.092896	89.590643	84.912281

compute the count of F durations and S durations, grouping by RType and LType.



LTypeSSum	RTypeFSum	RTypeSSum	RTypeFMean	RTypeSMean	LTypeFMean	LTypeSMean	RTypeFcount	RTypeScount	LTypeFcount	LTypeScount
14520.0	15020.0	14840.0	82.076503	81.092896	89.590643	84.912281	183	183	174	168

Once the sum, mean and counts for each of these changes are computed we have 48 variables computed. We will use these 48 variables as independent variables in the model.

Left Eye (LType)	Fixation	Sum / Mean / Count	3
Right Eye (RType)	Fixation	Sum / Mean / Count	3
Left Eye	Saccade	Sum / Mean / Count	3
Right Eye	Saccade	Sum / Mean / Count	3

Left Eye	Horizontal Movement	Left movement	Sum / Mean / Count	3
Left Eye	Horizontal Movement	Right Movement	Sum / Mean / Count	3
Left Eye	Horizontal Movement	No Change	Sum / Mean / Count	3
Left Eye	Vertical Movement	Left movement	Sum / Mean / Count	3
Left Eye	Vertical Movement	Right Movement	Sum / Mean / Count	3
Left Eye	Vertical Movement	No Change	Sum / Mean / Count	3
Right Eye	Horizontal Movement	Left movement	Sum / Mean / Count	3
Right Eye	Horizontal Movement	Right Movement	Sum / Mean / Count	3
Right Eye	Horizontal Movement	No Change	Sum / Mean / Count	3
Right Eye	Vertical Movement	Left movement	Sum / Mean / Count	3
Right Eye	Vertical Movement	Right Movement	Sum / Mean / Count	3
Right Eye	Vertical Movement	No Change	Sum / Mean / Count	3

So in total there are 12 + 36 + DistanceL and Distance R and Subject, Label and Gender = 53 variables, out of which 52 will be the independent variables and the label will be the dependent variable.

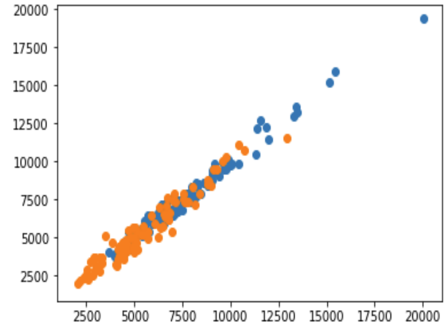
These features are extracted and saved as CSV for further analysis, for all 185 subjects – as one row per subject.

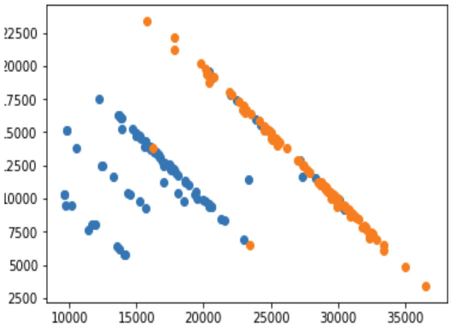
Storytelling and Inferential Statistics

The dataset consists of numerical values from the transformed features which has Gender, Label, Subject and 50 features derived from the movement of the eye and time that were recorded. Since it is derived from the data set there are no null or missing values. The information contained in these features captures different quantitative properties of eye movements in reading, including their duration, amplitude, direction, and stability

A simple set of low-level features were defined that ranged over both fixation and saccadic events. The eye movement analysis where the horizontal (LX, RX) and vertical (LY, RY) eye movement signal is plotted each other - fixations, and saccades.

Comparing the DistanceL and DistanceR on plot indicates it is similar between the controls and reading disability subjects. But the relationship LTypeFSum and LTypeSSum which denotes the fixation and saccade tells a different story. The fixation time differs for the subject and when it is plotted of the count with label it shows the complete distinction of the two.





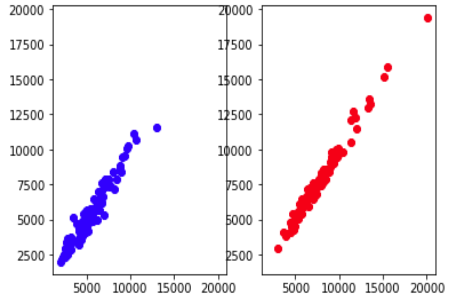
Compare the distance travelled by each eye for the control and dyslexic groups

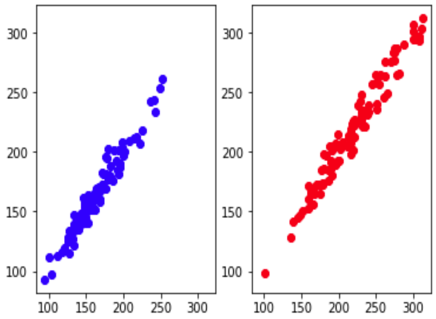
Looking at a scatter plot between the left and right eye movements - distance travelled, we see that both eyes travel almost equal distances, with some exceptions for dyslexic kids, their eyes traveling longer distances. Orange is control group and blue is dyslexic group.

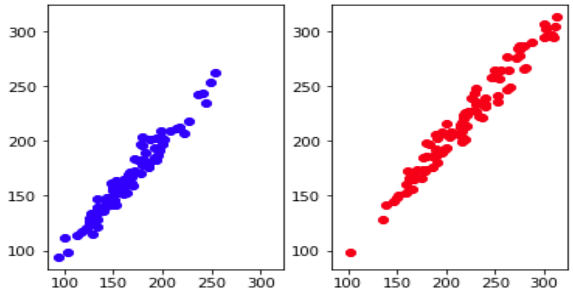
Then we look at the number of saccades and fixations for each of the groups. we see that the number of saccades as well as the fixations are almost the same between the right and left eyes - indicated by the slope of the line. However, the number of saccades and fixations are more for dyslexic kids than control kids.

This may be because dyslexic kids move their eyes more frequently than control groups to travel the same distance.

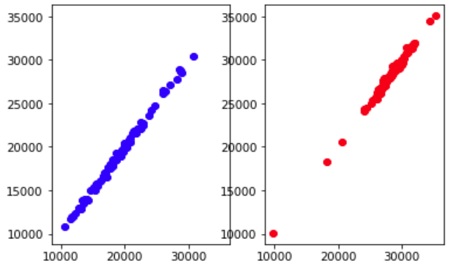
Note: these charts have the same scale for x and y axes

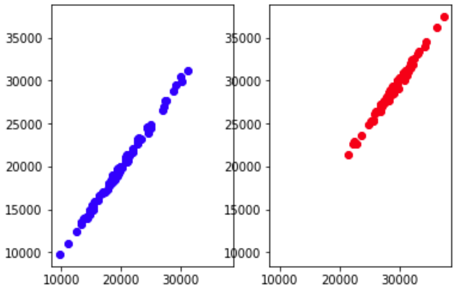

Compare left/right eye distance traveled for control and dyslexis groups

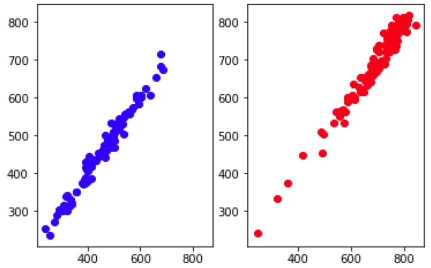

Compare the count of left/right eye fixations for control and dyslexis groups

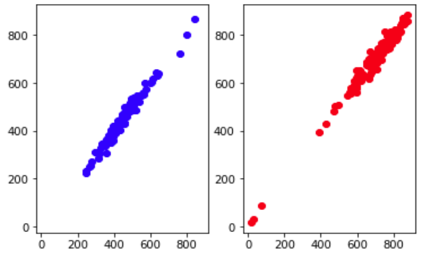

Compare the count of left/right eye saccades for control and dyslexis groups

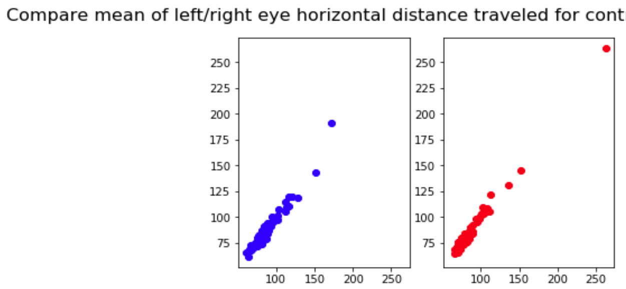
The sum of the horizontal/vertical distances moved by the dyslexic group as well as the count of horizontal/vertical distances moved by the dyslexic group are higher than the control group. But the mean distance traveled by the horizontal/vertical movements is about the same.

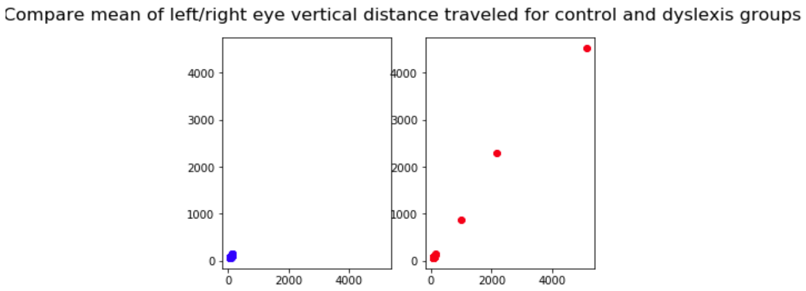

Compare left/right eye horizontal distance traveled for control and dyslexis groups


Compare left/right eye vertical distance traveled for control and dyslexis groups


Compare count of left/right eye horizontal movements for control and dyslexis groups

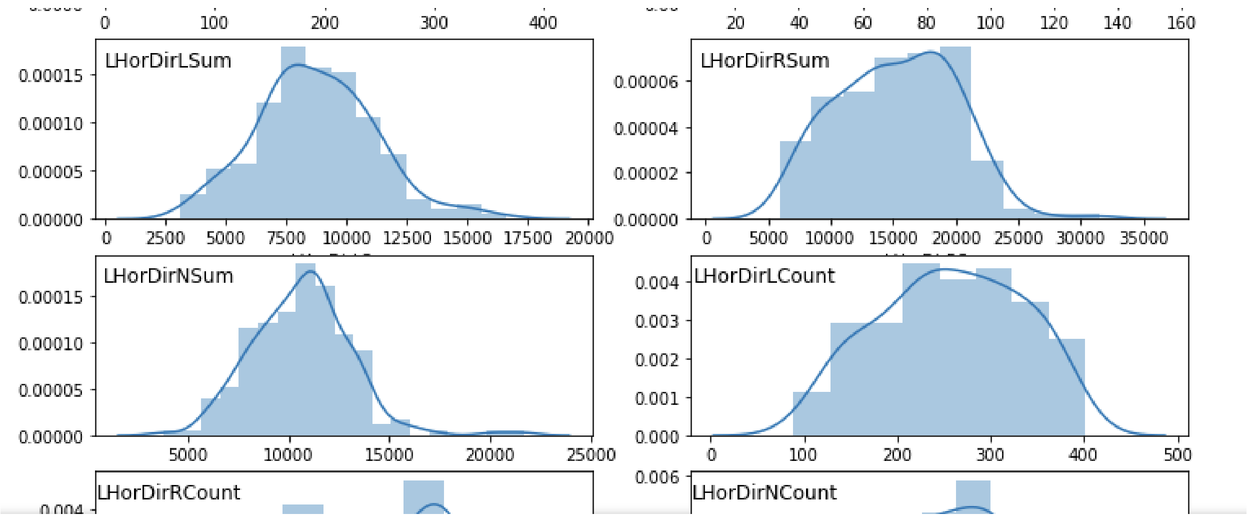

Compare count of left/right eye vertical movements for control and dyslexis groups



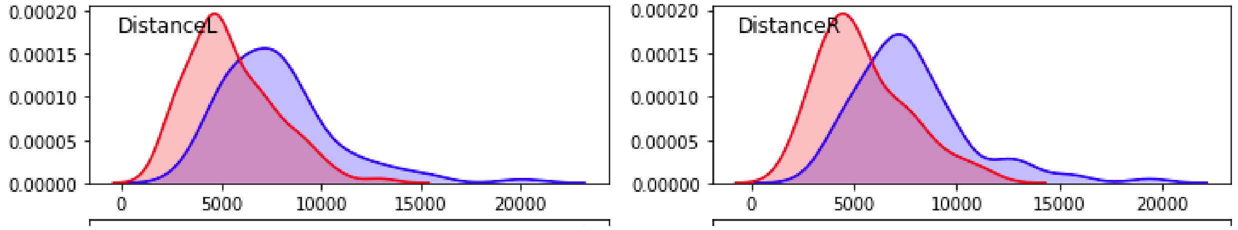


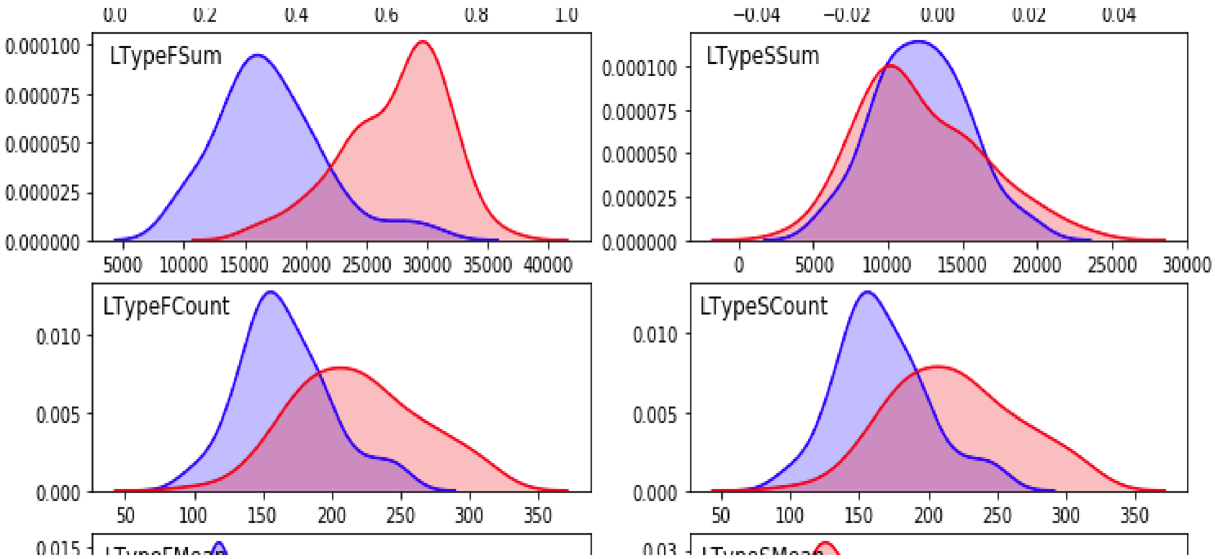
Observation of the scatter plot: the LHorDirLMean, LVerDirRMean, LVerDirUMean and LVerDirDMean-which represents the mean value of the eye movement in the horizontal and vertical, up and down direction seems to be very similar for the controls and high-risk subjects. Whereas the count and the sum of the same features shows that the cluster towards one value and not uniform.

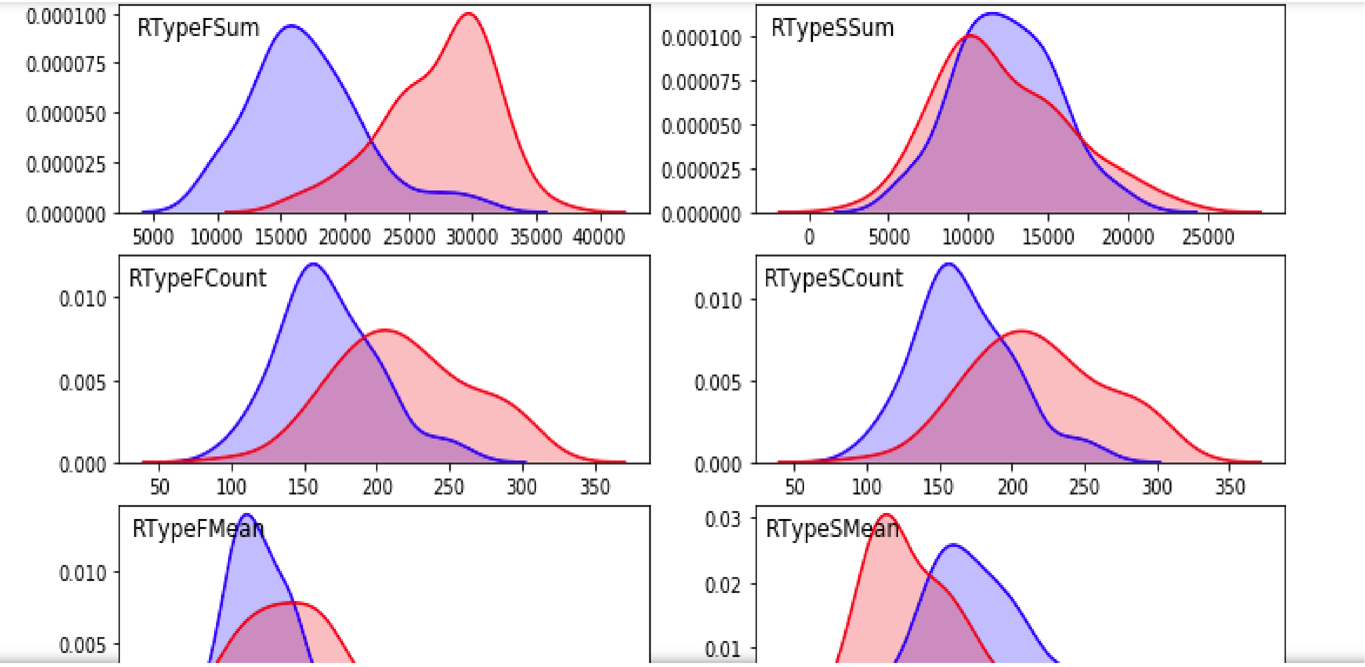
All the features are plotted in histogram and results indicates that variables such as LTypeFSum, LHorDirRSum, LHorDirLSum, RHorDirRCount etc have a major impact in determining the dyslexic nature



Let us visualize the probability distribution of both control and dyslexic groups in a single plot.




0.0
0.2
0.4
0.6
0.8
1.0
−0.04
−0.02
0.00
0.02
0.04
LTypeFSum
0.000100
0.000075
0.000050
0.000025
0.000000
5000
10000
15000
20000
25000
30000
35000
40000
LTypeSSum
0.000100
0.000075
0.000050
0.000025
0.000000
0
5000
10000
15000
20000
25000
30000
LTypeFCount
0.010
0.005
0.000
50
100
150
200
250
300
350
LTypeSCount
0.010
0.005
0.000
50
100
150
200
250
300
350
0.015
LTypeFMean
0.03
LTypeSMean


RTypeFSum
0.000100
0.000075
0.000050
0.000025
0.000000
5000
10000
15000
20000
25000
30000
35000
40000
RTypeSSum
0.000100
0.000075
0.000050
0.000025
0.000000
0
5000
10000
15000
20000
25000
RTypeFCount
0.010
0.005
0.000
50
100
150
200
250
300
350
RTypeSCount
0.010
0.005
0.000
50
100
150
200
250
300
350
RTypeFMean
0.010
0.005
RTypeSMean
0.03
0.02
0.01

Baseline Modeling

Dependent Variable

The column label is the prediction column. This is the dependent variable. 0 is control group and 1 is dyslexic group.

Independent Variables

There are 52 independent variables. All but gender are numerical/decimal variables. The gender column is a categorical variable and has 1 for males and 0 for females

Model Training Data

The dataset has 88 subjects in the control group and 97 subjects in the dyslexic group. This way the data seems to be very balanced. We use StandardScaler to scale the variables to a standard normal format.

A training and validation approach is used. The available data is split with 70% of the data used for training the model and 30% of the data used to validate the model as the test data set. The dependent and independent variables are split into X_train and y_train for the training data set and X_test and y_test for the testing dataset. A stratified ratio preserved rows are split across the test and training data.

The original data has a control/dyslexia ratio of (88/97) 90%, the training data has a ratio of (61/68) 90% and the test data has a ratio of (27/29) 93%.

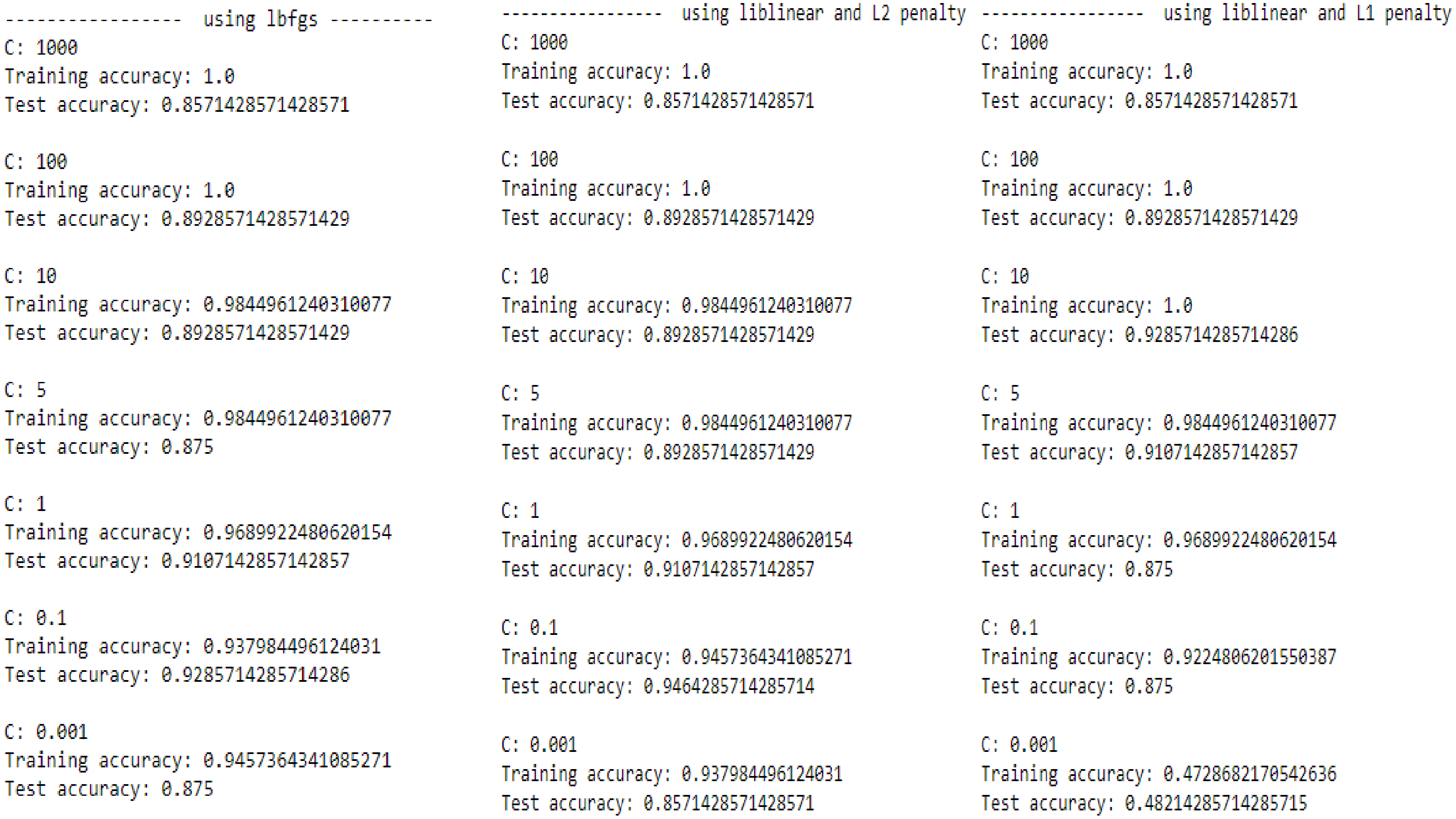
This is a binary classification problem, and we plan to use one of Logistic Regression or a Gradient Boosted Trees for modeling the data. Logistic regression is an appropriate regression analysis to conduct when the dependent variable is binary. The data analysis we did earlier confirm that there were no outliers in the data.

Mathematically, a logistic regression can be represented as

$$= \log\left(\frac{p(y=1)}{1-(p=1)}\right) = \beta_0 + \beta_1 \cdot x_{i2} + \beta_2 \cdot x_{i2} + \_ + \beta_p \cdot x_{in}$$

We try to apply the ski learn Logistic regression model and see if there were any overfitting. In addition, as a competing model, we also try using decision trees. We use a Gradient Boosting Tree model to make predictions on the same set of data.

Now to find the hyperparameters for a logistic regression model, we try with different C values, 1000 through 0.001. Looking at the results a C value of 10 provides the best training and test accuracy. We try different values of the learning rate C and different hyperparameters for the model (liblinear with L1 penalty, liblinear with L1 penalty, and lbgfs.



```
-----------------  using lbfgs ----------
C: 1000
Training accuracy: 1.0
Test accuracy: 0.8571428571428571

C: 100
Training accuracy: 1.0
Test accuracy: 0.8928571428571429

C: 10
Training accuracy: 0.9844961240310077
Test accuracy: 0.8928571428571429

C: 5
Training accuracy: 0.9844961240310077
Test accuracy: 0.875

C: 1
Training accuracy: 0.9689922480620154
Test accuracy: 0.9107142857142857

C: 0.1
Training accuracy: 0.937984496124031
Test accuracy: 0.9285714285714286

C: 0.001
Training accuracy: 0.9457364341085271
Test accuracy: 0.875
```

```
----------------  using liblinear and L2 penalty
C: 1000
Training accuracy: 1.0
Test accuracy: 0.8571428571428571

C: 100
Training accuracy: 1.0
Test accuracy: 0.8928571428571429

C: 10
Training accuracy: 0.9844961240310077
Test accuracy: 0.8928571428571429

C: 5
Training accuracy: 0.9844961240310077
Test accuracy: 0.8928571428571429

C: 1
Training accuracy: 0.9689922480620154
Test accuracy: 0.9107142857142857

C: 0.1
Training accuracy: 0.9457364341085271
Test accuracy: 0.9464285714285714

C: 0.001
Training accuracy: 0.937984496124031
Test accuracy: 0.8571428571428571
```

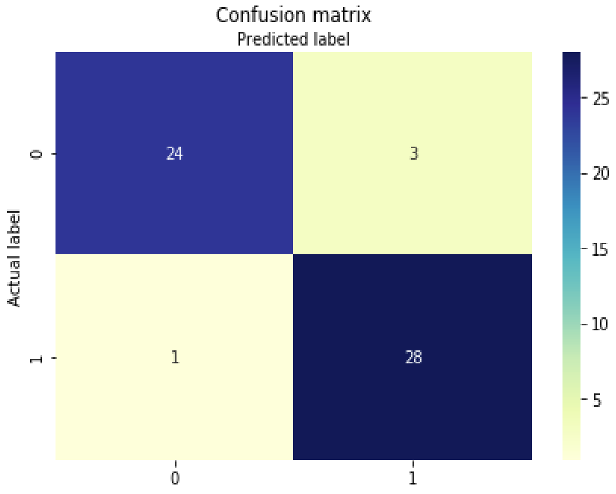
```
----------------  using liblinear and L1 penalty
C: 1000
Training accuracy: 1.0
Test accuracy: 0.8571428571428571

C: 100
Training accuracy: 1.0
Test accuracy: 0.8928571428571429

C: 10
Training accuracy: 1.0
Test accuracy: 0.9285714285714286

C: 5
Training accuracy: 0.9844961240310077
Test accuracy: 0.9107142857142857

C: 1
Training accuracy: 0.9689922480620154
Test accuracy: 0.875

C: 0.1
Training accuracy: 0.9224806201550387
Test accuracy: 0.875

C: 0.001
Training accuracy: 0.4728682170542636
Test accuracy: 0.48214285714285715
```

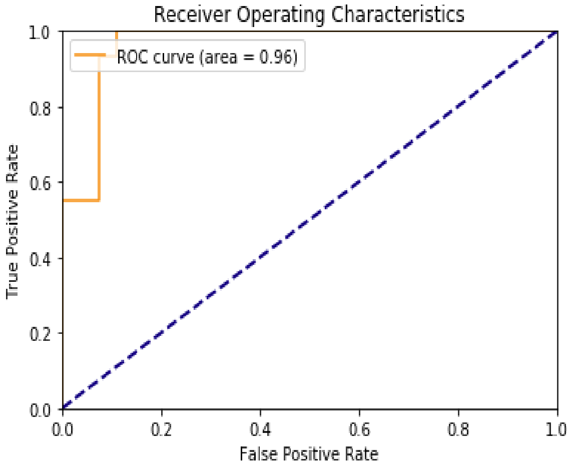
Of all these the best one seems to be a liblinear regression model with C=10. We try to fit a Logistic regression model with a regularization parameter (C) of 10 and a maximum of 500 iterations. Then test the model using predictions from the test data set.

We prepare and print the classification report for the training and test data. The ROC looks reasonably good for an initial model.

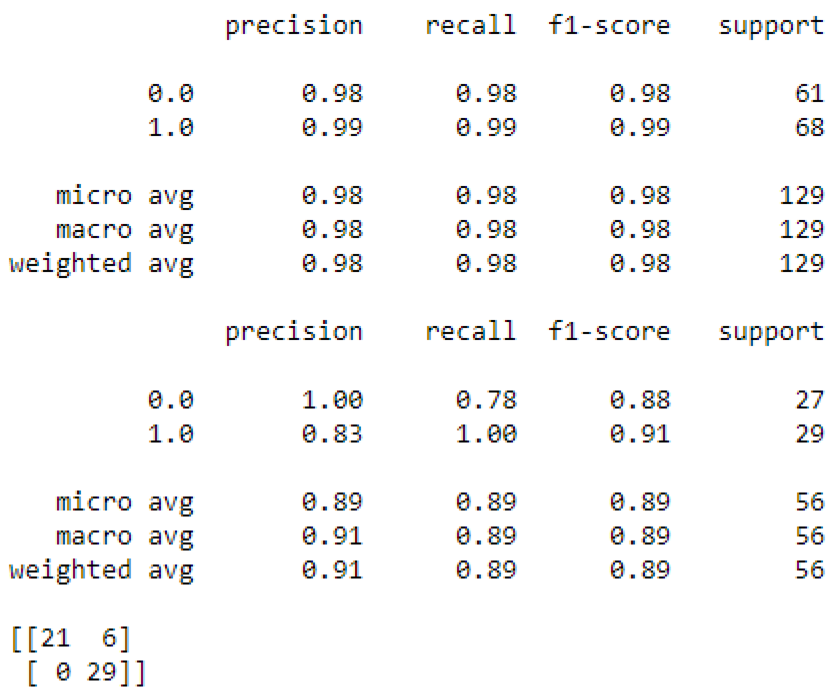
model = LogisticRegression (penalty='l1'

```
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00        61
         1.0       1.00      1.00      1.00        68

   micro avg       1.00      1.00      1.00       129
   macro avg       1.00      1.00      1.00       129
weighted avg       1.00      1.00      1.00       129
              precision    recall  f1-score   support

         0.0       0.96      0.89      0.92        27
         1.0       0.90      0.97      0.93        29

   micro avg       0.93      0.93      0.93        56
   macro avg       0.93      0.93      0.93        56
weighted avg       0.93      0.93      0.93        56

[[24  3]
 [ 1 28]]
```

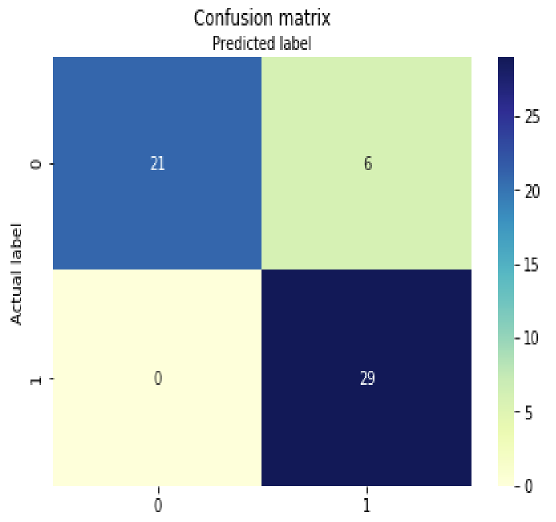


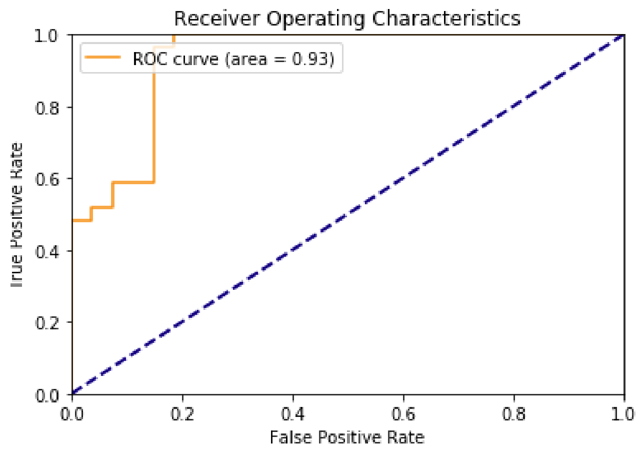


for the model = LogisticRegression (solver='lbfgs'



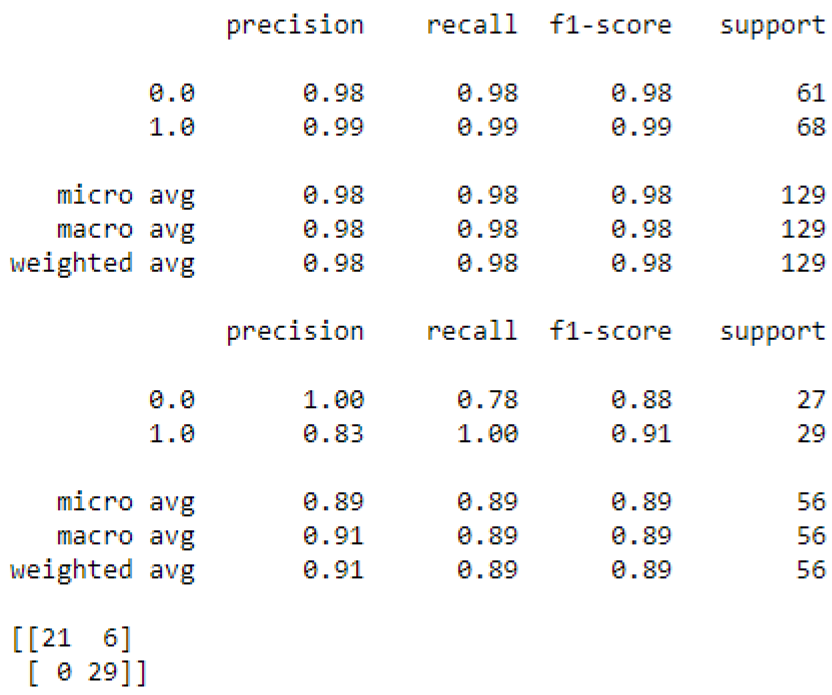
```
              precision    recall  f1-score   support

         0.0       0.98      0.98      0.98        61
         1.0       0.99      0.99      0.99        68

   micro avg       0.98      0.98      0.98       129
   macro avg       0.98      0.98      0.98       129
weighted avg       0.98      0.98      0.98       129
              precision    recall  f1-score   support

         0.0       1.00      0.78      0.88        27
         1.0       0.83      1.00      0.91        29

   micro avg       0.89      0.89      0.89        56
   macro avg       0.91      0.89      0.89        56
weighted avg       0.91      0.89      0.89        56

[[21  6]
 [ 0 29]]
```



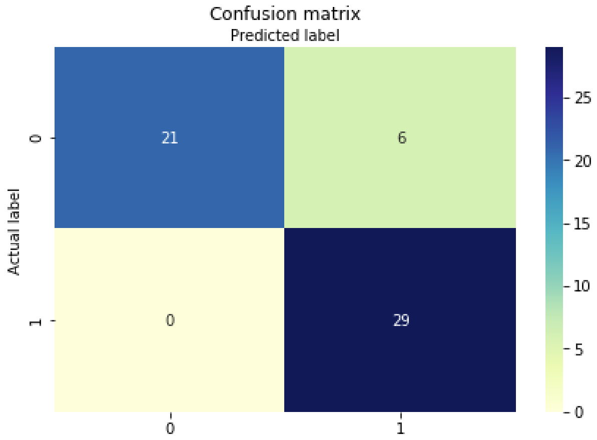


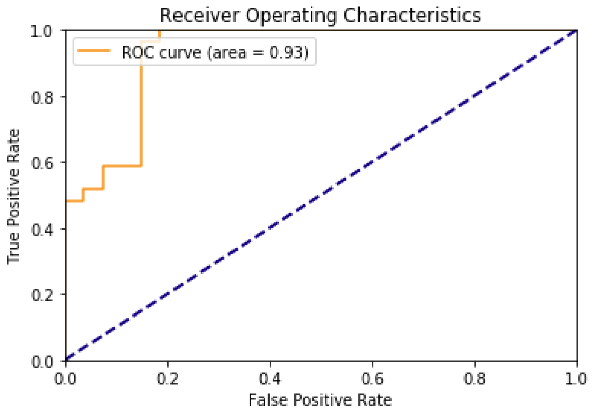
Extended Modeling

Now trying with the Gradient Boosting Classifier, to see if there are any better results.



```
              precision    recall  f1-score   support

         0.0       0.98      0.98      0.98        61
         1.0       0.99      0.99      0.99        68

   micro avg       0.98      0.98      0.98       129
   macro avg       0.98      0.98      0.98       129
weighted avg       0.98      0.98      0.98       129

              precision    recall  f1-score   support

         0.0       1.00      0.78      0.88        27
         1.0       0.83      1.00      0.91        29

   micro avg       0.89      0.89      0.89        56
   macro avg       0.91      0.89      0.89        56
weighted avg       0.91      0.89      0.89        56

[[21  6]
 [ 0 29]]
```





It looks like the linear regression with solver = liblinear and L1 penalty provides the best results

Conclusions and Future Work

Children with dyslexia regularly spend many years struggling in school before receiving suitable specialized support. Useful screening methods can be implemented in school settings to respond to this situation and enable earlier support.

Study of the use of eye tracking during reading as a screening method helps to produce individual-level predictions with high sensitivity and specificity in less than a minute of tracking time. The only response measured was the eye movement signal and since that itself is objective it is neither right nor wrong according to some predefined criteria. Also, a screening test based on eye tracking may reduce the amount of stress that more traditional test methods inflict, since subjects may be more likely to experience that they are engaged in a task by themselves rather than explicitly performing a task for someone else.

Future Work

Early identification of individuals in need of support is the first important step in this process. For this purpose, using eye tracking during reading may prove particularly useful.

Recommendations for the Clients

Based on my analysis, I would recommend the following:

1. It is recommended to use the logistic regression with C=10, solver= liblinear with L1 penalty to develop the model

Consulted Resources

1. https://figshare.com/collections/Screening_for_Dyslexia_Using_Eye_Tracking_During_Reading/3521379/1
2. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0165508#sec002